CMPSC 465 Data Structures & Algorithms
Spring 2024 Mehrdad Mahdavi and David Koslicki

HW 9

Due April 4, 10:00 pm

Instructions: You are encouraged to solve the problem sets on your own, or in groups of up to five people, but you must write your solutions strictly by yourself. You must explicitly acknowledge in your write-up all your collaborators, as well as any books, papers, web pages, etc. you got ideas from.

Formatting: Each problem should begin on a new page. Each page should be clearly labeled with the problem number. The pages of your homework submissions must be in order. You risk receiving no credit for it if you do not adhere to these guidelines.

Late homework will not be accepted. Please, do not ask for extensions since we will provide solutions shortly after the due date. Remember that we will drop your lowest two scores.

You need to submit it via Gradescope (Class code XX7RVV). Please ask on Campuswire about any details concerning Gradescope and formatting.

For each algorithm question, explain your algorithm and analyze its correctness and running time. Pseudocode is not required, but you may include it if you feel it makes your written explanation more clear.

1. (20 pts.) **Set Cover.** Consider each of the following words as a set of letters (e.g. treat the word "ash" as $S_1 = \{a, s, h\}$): {dog, goat, ate, dot, timer, road, tread, eye}. Which sets are selected by the greedy heuristic and in what order? When breaking ties, pick the word that appears first in the list of words. Show the number of uncovered elements in the chosen set for each iteration.

2. (20 pts.) **Huffman Encoding.** We use Huffman's algorithm to obtain an encoding of alphabet $\{a, b, c\}$ with frequencies f_a, f_b, f_c. In each of the following cases, either give an example of a specific set of frequencies f_a, f_b, f_c that would yield the specified encoding or explain why the code cannot possibly be obtained.

(a) $\{0, 10, 11\}$

(b) $\{0, 00, 10\}$

(c) $\{00, 10, 11\}$

3. (20 pts.) **Huffman Properties.** Prove the following two properties of the Huffman encoding scheme.

(a) If some character occurs with frequency more than $2/5$, then there is guaranteed to be a codeword of length 1.

(b) If all characters occur with frequency less than $1/3$, then there is guaranteed to be no codeword of length 1.

4. (20 pts.) **Colored Number Line.** You are given n points on a number line. Each point has a color; red, green, or blue. You can draw an undirected edge from one point to any other point, with the cost of the edge as the distance between the two points. Write an efficient algorithm that outputs the minimal cost of edges such that if all the blue points are removed (associated edges are removed), the remaining points are all connected (there is a path between them); the same thing applying if all the red points are removed. Explain why your algorithm works and its running time.

5. (20 pts.) **Huffman Efficiency.** We will consider how much Huffman encoding can compress a file F of m characters taken from an alphabet of $n = 2^k$ characters $x_0, x_1, \ldots, x_{n-1}$ (each character appears at least once).

(a) Let $S(F)$ represent the number of bits it takes to store F without using Huffman coding (i.e., using the same number of bits for each character). Find an expression (formula) for $S(F)$ in terms of m, n, and/or k.

(b) Let $H(F)$ represent the number of bits used in the optimal Huffman coding of F. We define the *efficiency* $E(F)$ of a Huffman coding on F as $E(F) := S(F)/H(F)$. For each m and n, give the minimum value of $E(F)$ and describe a file F for which $E(F)$ is as small as possible.

(c) For each m and n describe a file F for which $E(F)$ is as large as possible (remember that each character must appear at least once). How does the largest possible efficiency increase as a function of n? Assume the file is much larger than the alphabet such that $n \log n = O(m)$. Give your answer in big-O notation as a function of n and/or k.

6. (0 pts.) **Acknowledgments.** The assignment will receive a 0 if this question is not answered.
(a) If you worked in a group, list the members of the group. Otherwise, write "I did not work in a group."
(b) If you received significant ideas about the HW solutions from anyone not in your group, list their names here. Otherwise, write "I did not consult with anyone other than my group members."
(c) List any resources besides the course material that you consulted in order to solve the material. If you did not consult anything, write "I did not consult any non-class materials."